Comparison of chest radiograph reading methods for assessing progress of pneumoconiosis over 10 years in Wittenoom crocidolite workers

N H de Klerk, A W Musk, A James, J J Glancy, W O C M Cookson

Abstract

Thirty three pairs of chest radiographs taken up to 10 years apart were obtained for 33 subjects suffering from asbestosis who had applied for compensation to the Pneumoconiosis Medical Board of Western Australia. Multiple films from the period before the first radiograph in each pair, from the intervening period between the two, and from the period subsequent to the second radiograph were also available and all films were read by two independent readers according to the 1980 ILO classification of pneumoconiosis. Films were read twice as side by side pairs ten years apart, twice as two separate randomly ordered films ten years apart, and once as part of the full series of all available chest radiographs on each subject to assess which method provided the best consistency (between reader variation) and repeatability (within reader variation). Judging by consistency, the full series method performed as well as either of the other methods when assessing radiographic changes and significantly better when assessing the level of profusion of small opacities. There was little to choose between the other two methods either judging by consistency or repeatability, which could not be estimated for the full series method. Use of all available films for a subject is recommended for assessing single films, as in a prevalence study, as well as for documenting change in a longitudinal study.

NH & MRC Unit of Epidemiology and Preventive Medicine, Department of Medicine, University of Western Australia, Nedlands, Western Australia 6009
N H de Klerk
Department of Respiratory Medicine, Sir Charles Gairdner Hospital, Perth
A W Musk, A James, W O C M Cookson
Department of Diagnostic Radiology, Sir Charles Gairdner Hospital, Perth
J J Glancy

The ILO/UICC classification of radiographs for the pneumoconioses[1] is well established for describing and quantifying radiograph appearances in subjects occupationally exposed to asbestos and other dusts.[2–4] Using this system, the decision on which strategy of reading to use when examining large series of chest radiographs for the presence and progression of pneumoconiosis will depend on the objectives of the study being performed and on which method will provide the most valid information.

It has been suggested that if more than one film is available use of the additional film(s) may bias the estimate of severity of the film of interest, implying that all films should be read independently and in random order.[5] This question does not arise in a prevalence survey if there is only one film available for each person and independent randomised reading and scoring to the ILO/UICC classification is the only possibility.

In assessing progression of radiographic pneumoconiosis from serial films there are several choices available. The use of side by side reading to the ILO/UICC classification irrespective of the particular objectives of the study is one compromise that does not meet the objectives of all studies because of the possible bias referred to above. Direct progression scoring (DPS) alone does not provide prevalence data as it does not include ILO/UICC scoring at the time of determining progression.[6] Independent random readings of all radiographs in a series, taken singly or in pairs, appears much more likely to demonstrate regression of changes[6] and has been shown to result in larger estimates of progression than direct progression scoring in one study.[5] Both methods, however, may suffer as a result of loss of the information that additional radiographs could provide in interpreting the film of interest especially if the quality of the radiographs is inconsistent.

Since there is no independent and absolute standard by which to judge the true degree of pulmonary fibrosis on any occasion there is no acceptable method for testing or measuring the validity of the strategy of assessing radiographic abnormality or change. Therefore this study was designed to assess

which method is superior as judged by consistency (inter-reader agreement) and repeatability (intra-reader agreement) in subjects who had made a compensation claim to the Pneumoconiosis Medical Board (PMB) of Western Australia (WA) for asbestosis, who had been exposed to crocidolite from Wittenoom Gorge, WA, and who had serial chest radiographs covering a period of at least 10 years.

Subjects

Thirty three of the 384 men who had applied for compensation, consisting of all those who had been awarded some compensation for asbestosis by the Pneumoconiosis Medical Board of Western Australia and for whom at least one pair of plain chest radiographs taken 10 years apart after the start of employment could be located, were included in this study. In addition, each man had at least one film taken during the periods before, between, and subsequent to the times of the relevant pair. Characteristics of these 33 men are given in table 1.

Methods

Thirty three pairs of chest radiographs, taken 10 years apart, for the 33 men were obtained from the Perth Chest Clinic and Perth teaching hospitals. Between six and 21 additional films, with an average of 13, taken during the previous, intervening, and subsequent periods for the particular pair of films were available for each man.

The films were read five times on different occasions over 12 months by two independent readers using the 1980 ILO classification of radiographs of the pneumoconioses[1] with the following three methods or strategies:

Pairs method—The films in each pair were placed side by side and viewed and scored simultaneously. The temporal order of the films was known.

Random singles method—The films in each pair were read separately in random order.

Full series methods—The films in each pair were placed in series with all other available films for that man and viewed simultaneously.

The first two methods were repeated by each reader on two separate occasions. Each reader, however, examined the full series simultaneously only once.

Table 1 Characteristics of the men

	Median	*Range*
Year of starting work	1955	1947–65
Year of birth	1925	1909–43
Age starting work	29	18–54
Duration of employment (days)	1430	70–6003
Average intensity of exposure (fibres/ml)	20	5–110
Total cumulative exposure (fibres/ml years)	85	3·8–1808
Workplace	*No of men*	
Mill only	7	
Mine only	14	
Mill and mine	5	
Elsewhere	7	
Total	33	

Table 2 Sample tables of agreement between and within readers

(a) *Profusion of small opacities (major ILO categories) on second film*

Random singles method (2nd occasion)

		Reader 2			
		*0**	*1*	*2*	*3*
Reader 1	0*	0	7	0	0
	1	0	9	6	0
	2	0	3	4	2
	3	0	0	0	2

Full series method

		Reader 2			
		*0**	*1*	*2*	*3*
Reader 1	0*	2	1	0	0
	1	0	14	1	0
	2	0	2	9	3
	3	0	0	0	1

(b) *Ten year changes in profusion*

Pairs method (reader 1)

		Occasion 2			
		1†	*2*	*3*	*4*
Occasion 1	1†	4	2	0	0
	2	0	15	1	0
	3	0	2	4	1
	4	0	0	3	1

Pairs method (reader 2)

		Occasion 2			
		1†	*2*	*3*	*4*
Occasion 1	1†	1	2	1	0
	2	0	2	4	1
	3	2	3	5	1
	4	0	0	0	1

*Major ILO categories.
†Changes grouped as: 1 No change or regression, 2 One or two minor ILO categories. 3 Three or four minor ILO categories, and 4 Five or more minor ILO categories.

The changes over ten years were classified into four groups: (1) No change or regression, (2) one or two minor ILO categories, (3) three or four minor ILO categories, and (4) five or more minor ILO categories.

Performance achieved by the different methods was assessed by use of log linear models fitted to the cross classification tables of both inter-reader agreement (consistency) and intra-reader agreement (repeatability) following the guidelines suggested by Tanner and Young[7] and using the computer program GLIM.[8] Table 2 shows some examples of the cross classification tables used: firstly, the agreement between reader 1 and reader 2 using the random singles method on the second occasion and the reader agreement using the full series method and, secondly, within reader agreement on assessment of change using the pairs method for both readers.

To ensure independence of the observations only the later of each pair of radiographs was considered for analysis either in terms of the level of radiographic abnormality on that radiograph or the change from the one 10 years earlier.

Tanner and Young discuss six basic models which may be applied to the two way cross classification of ordered scores given by two observers (extensions to more than two observers or analyses by occasions instead of observers are straightforward[7]):

(1) *No association*—The classification only depends on the rows and columns—analogous to the usual chi-squared test.
(2) *Homogeneous disagreement*—All observations are not on the main diagonal of perfect agreement but the spread is uniform over the other cells—useful for nominal data and analogous to the kappa statistic.[9]
(3) *Systematic direction bias* is similar to homogeneous disagreement but the amount of disagreement is different either side of the main diagonal—that is, one observer measures "higher" than the other.
(4) *Symmetric band disagreement* is also similar to homogeneous disagreement but the amount of disagreement varies with the level of disagreement analogous to the weighted kappa statistic[9] giving obvious advantages over the kappa statistic when examining ordinal data of this type.
(5) *Asymmetric band disagreement*—A combination of systematic direction bias and symmetric band disagreement with different levels for each observer.
(6) *Symmetric cell disagreement*—Similar to symmetric band disagreement but the symmetry only applies to cells—that is, the disagreement depends on the level of measurement as well as the level of disagreement.

It was not possible to examine models (5) or (6) given the amount of data in the study, and model (1) (no association) was not relevant.

The equation for the first fitted model (homogeneous disagreement) is:

$$\log m_{ij} = \text{const} + a_i + b_j + c.d_{ij}$$

where m is the number of observations in the ith row and the jth column, a and b are the respective row and column coefficients, and c is the coefficient for a variable d_{ij} which = 0 if i = j and = 1 otherwise. It can be seen to be equivalent to the use of the kappa statistic[9] by, firstly, allowing for likely chance agreement with the row and column coefficients and, secondly, allocating equal weight to all disagreements. It is thus most applicable to nominally scaled data but is still an improvement on the percentage agreement which makes no allowance for unequal proportions in the different groups.

The equation for the second fitted model (systematic direction bias) is:

$$\log m_{ij} = \text{const} + a_i + b_j + c.d_{ij}$$

the same as the previous equation except that c is the coefficient for a variable d_{ij} which equals 0 if $i = j$, 1 if $i < j$ and 2 if $i > j$, thus allowing for unidirectional bias in either of the observers.

The equation for the third fitted model (symmetric band disagreement) is:

$$\log m_{ij} = \text{const} + a_i + b_j + c.d_{ij}$$

the same as the previous equation except that c is the coefficient for a variable d_{ij} equalling the absolute value of the difference between i and j and is equivalent to the use of a weighted kappa statistic with arbitrary weights for the level of disagreement.[9] Incorporating such information on the magnitude of the disagreement is an essential property of a measure of agreement on an ordinal scale, and appears to be an improvement on methods used previously for this type of data.[10]

For all three models the resulting coefficient estimates, when exponentiated, may be interpreted as relative disagreement rates. Thus a value of 1·0 implies that observers are equally likely to agree as not to agree (at a particular level) and the level of agreement is only what would be expected by chance (equivalent to a kappa of zero). A value of 0·5 implies that observers are twice as likely to agree as not to agree and a value of 0·1 implies that observers are ten times as likely to agree as not to agree—that is, the smaller the coefficient the better the agreement. Goodnesss of fit of the different models was assessed by comparison of the residual deviance with the appropriate chi-squared distribution and the statistical significance of additions to each model was estimated by the corresponding difference in residual deviance.[8]

There were four sets of tables to which these models were applied. The two sets of tables of agreement between and within readers on the ILO classification of the second film in each pair, and the two sets of tables of agreement between and within readers on the change from the first film to the second film graded as described above. After fitting rows, columns, and each of the above disagreement terms separately, the same additional procedure was followed. Firstly, the interaction between method of reading and the particular disagreement term was added, then the interaction between the disagreement term with reader (for repeatability analyses) or occasion (for consistency analyses), and, finally, the interaction of these terms with the method used was added. These latter effects were used to assess reader quality. A further assessment of reader quality could be made within this same framework by comparing the same reader across different methods. This was not done here because the study aimed to compare methods, not readers. Kendall's tau correlation coefficient was also calculated for the table for each method[11] averaging across occasions for the consistency tables and averaging across readers for the repeatability tables.

Results

SINGLE RADIOGRAPHS

For all comparisons the goodness of fit of the first model (homogeneous disagreement) and the second model (systematic direction bias) was poor ($p < 0.01$) whereas the fit of the third (symmetric band disagreement) was always reasonable ($p > 0.10$). Addition of a term for systematic direction bias to the symmetric band model had little

Table 3 Assessment of single radiographs

(a) *Consistency (inter-reader agreement) in assessing single radiographs*	*Pairs*	*Random singles*	*Full series*	*p Value**
Relative disagreement rates (95% CI):				
Homogeneous (all categories)	0·27 (0·16–0·47)	0·31 (0·19–0·52)	0·14 (0·06–0·33)	0·04
Symmetric band ± 1 category	0·30 (0·18–0·53)	0·36 (0·21–0·62)	0·16 (0·07–0·38)	0·02
± 2, 3 categories	0·01 (0·00–0·07)	0† —	0† —	
Kendall's tau correlation coefficient	0·65	0·66	0·80	

(b) *Repeatability (intra-reader agreement) in assessing single radiographs*	*Pairs*	*Random singles*	*p Value**
Relative disagreement rates (95% CI):			
Homogeneous (all categories)	0·28 (0·17–0·47)	0·20 (0·12–0·36)	0·24
Symmetric band ± 1 category	0·40 (0·24–0·66)	0·29 (0·17–0·51)	0·36
± 2,3 categories	0†	0†	
Kendall's tau correlation coefficient	0·63	0·69	

*p Value for testing equality of relative rates for all methods.
†No disagreement at this level.

effect in any of the four sets ($p > 0·8$ in all cases).

Table 3(a) shows the relative disagreement rates for the first set of tables, that of agreement between readers on the level of profusion of small opacities on the final film of each pair. The first two cross classifications in table 2 are examples of the tables used. Whereas the pairs method and the random singles method are equally good clearly the full series method performs better than either of them with a relative disagreement rate 0·16 (95% CI 0·07–0·38) for disagreement by one major category and no disagreement of two or more. The better performance of the full series method appeared statistically significant given the significant test of heterogeneity of the symmetric band terms over the different methods. Addition of interaction terms between method and occasion of reading were not significant ($p > 0·2$).

With respect to repeatability (table 3b), the random singles method appeared to perform better than the pairs method with a relative disagreement rate of 0·29 compared with 0·4 for disagreement by one major category, although the difference could have been due to chance ($p = 0·4$). Addition of interaction terms between readers and methods were not significant either ($p > 0·6$). Interestingly, the pairs method showed better agreement between readers than within, whereas the opposite was true for the random singles method.

RADIOGRAPHIC CHANGES

The rates of disagreement were higher for all meth-

Table 4 Assessment of radiographic changes

(a) *Consistency (inter-reader agreement) in assessing radiographic changes over 10 years*	*Pairs*	*Random singles*	*Full series*	*p Value**
Relative disagreement rates (95% CI):				
Homogeneous (all categories)	0·45 (0·28–0·74)	0·44 (0·27–0·72)	0·39 (0·21–0·72)	0·77
Symmetric band ± 1 category	0·58 (0·35–0·97)	0·56 (0·33–0·94)	0·54 (0·28–1·03)	0·92
± 2,3 categories	0·21 (0·09–0·48)	0·21 (0·09–0·48)	0·12 (0·03–0·49)	
Kendall's tau correlation coefficient	0·35	0·45	0·63	

(b) *Repeatability (intra-reader agreement) in assessing radiographic changes over 10 years*	*Pairs*	*Random singles*	*p Value**
Relative disagreement rates (95% CI):			
Homogeneous (all categories)	0·30 (0·18–0·49)	0·35 (0·22–0·56)	0·38
Symmetric band ± 1 category	0·37 (0·22–0·63)	0·49 (0·30–0·80)	0·43
± 2,3 categories	0·15 (0·06–0·36)	0·10 (0·04–0·28)	
Kendall's tau correlation coefficient	0·48	0·53	

*p Value for testing equality of relative rates for all methods.

ods of evaluating radiographic changes (table 4a). The relative disagreement rates for consistency were similar for all three methods. Repeatability when assessing changes was not much worse than when assessing level (table 4b). There was, however, a significant interaction between reader and method ($p = 0{\cdot}01$). The second reader had a much higher rate of disagreement than reader 1 using the pairs method or than either reader 1 or himself using the random singles method. Part of the difference may be seen in the lower two cross-tabulations in table 2.

Kendall's tau correlation coefficient showed a similar pattern being worst (lowest) for between reader agreement on changes, slightly better for within reader agreement on changes, better still for both between and within reader agreement on profusion level using the pairs and random singles methods, and best of all for between reader agreement in assessing profusion level using the full series method. Agreement between readers was reasonably close to and comparable with other studies which used Kendall's tau to estimate agreement.[3 4]

Discussion

Using the criterion of consistency between readers, this study has shown the superiority of using all available films for each person when assessing the degree of abnormality of single films (as in a prevalence study). It has also shown that consistency in assessing the amount of progression in ILO grades of abnormality between films over 10 years is at least as good using this method as it is when grading the films independently or just as a pair. There appeared to be little or no difference in consistency or repeatability when scoring films as ordered pairs simultaneously or as single films separately.

Liddell and Morgan have reviewed the methods of reading films for assessing progression and recommend side by side reading of pairs because it is simple and as good as other methods.[6] The results of this study do not contradict this recommendation for evaluating progression, especially when time savings are considered. Direct progression scoring has been examined by McMillan *et al* with the suggestion that using the full series together with the ILO classification may bias the differences.[5] There is certainly no evidence of this bias here unless it is the same for both readers. Direct progression scoring suffers from the drawback that no ILO grade is recorded unless it is carried out as a separate exercise and the method yields less progression of small opacities than random reading of single films. It is thus (surprisingly) a less sensitive method for recognising progression.

Side by side reading with disguise of the temporal order of the films has also been attempted but in the current study, as in most, the older films were easy to recognise.

The results of this study are therefore in agreement with the previous studies with regard to assessing progression but clearly show the superiority of the full series method when scoring individual radiographs. Given that most studies of radiographic change will also need to include assessment of the level of profusion in individual radiographs it is concluded that, where possible, the inclusion of greater numbers of films in the procedure is recommended.

1 International Labour Office. *Guidelines for the use of ILO international classification of radiographs of pneumoconioses.* Geneva: ILO, 1980. (Occupational safety and health series, No 22.)
2 Liddell FDK. Validation of the UICC/Cincinnati classification of radiographs in terms of prediction of mortality of asbestos workers. In: Wagner JC, ed. *Biological effects of mineral fibres.* Lyon: IARC, 1980:667–7. (Sci publ No 30.)
3 Musk AW, de Klerk NH, Cookson WOCM, Morgan WKC. Radiographic abnormalities and duration of employment in Western Australian iron-ore miners. *Med J Aust* 1988;**148**:332–4.
4 Cookson WOCM, de Klerk NH, Musk AW, Armstrong BK, Glancy JJ, Hobbs MST. The prevalence of radiographic asbestosis in crocidolite miners and millers at Wittenoom, Western Australia. *Br J Ind Med* 1986;**43**:450–7.
5 McMillan GHG, Rossiter CE, Deacon R. Comparison of independent randomised reading of radiographs with direct progression scoring for assessing change in asbestos related pulmonary and pleural lesions. *Br J Ind Med* 1982;**39**:60–1.
6 Liddell FDK, Morgan WKC. Methods of assessing serial films of the penumoconioses; a review. *J Soc Occup Med* 1978;**28**: 6–15.
7 Tanner MA, Young MA. Modelling ordinal scale disagreement. *Psychol Bull* 1985;**98**:408–15.
8 Baker RJ, Nelder JA. *The GLIM system release 3.* Oxford: Numerical Algorithms Group, 1978.
9 Fleiss JL. *Statistical methods for rates and proportions.* 2nd ed. New York: Wiley, 1981:212–36.
10 Musch DC, Landis JR, Higgins ITT, Gilson JC, Jones RN. An application of kappa-type analyses to interobserver variation in classifying chest radiographs for pneumoconiosis. *Stat Med* 1984;**3**:73–83.
11 Armitage P. *Statistical methods in medical research.* New York: Wiley, 1971:404.

Accepted 3 April 1989